ISSN 2519-2558
РОЗДІЛ 1. ПРОБЛЕМИ ЛІНГВІСТИКИ ТЕКСТУ ТА ДИСКУРСУ

Отримано: 21 листопада 2018 р.

Прорецензовано: 20 грудня 2018 р.

Прийнято до друку: 21 грудня 2018 р.

e-mail: lesya.kotsyuk@oa.edu.ua

yuriy.kotsyuk@oa.edu.ua

DOI: 10.25264/2519-2558-2018-4(72)-18-22

Коцюк Л. М., Коцюк Ю. А. Парадигма типологічних характеристик корпусу за типом текстових даних. *Наукові записки Національного університету «Острозька академія»: серія «Філологія»*. Острог : Вид-во НаУОА, 2018. Вип. 4(72), грудень. С. 18–22.

УДК 81'33

Коцюк Леся Миколаївна,
кандидат філологічних наук, доцент, Національний університет «Острозька академія»
Коцюк Юрій Анатолійович,
кандидат психологічних наук, ст. викладач, Національний університет «Острозька академія»

ПАРАДИГМА ТИПОЛОГІЧНИХ ХАРАКТЕРИСТИК КОРПУСУ ЗА ТИПОМ ТЕКСТОВИХ ДАНИХ

У статті робиться спроба проаналізувати типологічні характеристики корпусів текстів. Здійснено класифікацію корпусів з огляду на те, які текстові дані бралися до уваги при його укладанні, зокрема за ступенем їх спеціалізації, формальною природою та за мовним параметром. Виявлено, що парадигму за параметром «ступінь спеціалізації текстових даних» складають загальномовні та спеціалізовані корпуси. У свою чергу, у групі спеціалізованих корпусів тип текстових даних, які визначають назву корпусу, до якого вони входять та слугують параметром відбору, може визначатися жанровою, стилістичною, часовою, антропоцентричною, професійною, комунікативною, географічною чи соціальною природою мовної різноманітності. Також представлено приклади згаданих типів корпусів текстів. У статті представлено термінологічні еквіваленти назв корпусів за типом мовних даних в українській та англійській мовах.

Ключові слова: *загальномовний корпус, спеціальний корпус, текстові дані, тип корпусу, типологічні характеристики.*

Lesia Kotsiuk,
Cand.Sc. (Philology), Associate Professor, Ostroh Academy National University
Yurii Kotsiuk,
Cand.Sc. (Psychology), senior lecturer, Ostroh Academy National University

PARADIGM OF CORPUS TYPOLOGICAL CHARACTERISTICS BY THE TYPE OF TEXT DATA

The article attempts to analyze the typological characteristics of text corpora. The author proposes to classify corpora with consideration of different aspects of this modern linguistic notion, namely the nature of the text data included in the corpus, the design and structural features of the corpus, the method of fixing and indexing text data in the corpus, as well as the way of how the corpus can be used. Particular attention in the article is paid to the classification of corpora considering what text data was taken into account while it was compiled, in particular by the degree of text data specialization, its formal nature, and the language parameter. It is revealed that the paradigm of the «degree of specialization of text data» is composed of general and specialized corpora. In its turn, inside the group of specialized corpora, the type of text data that defines the name of the corpus to which they belong and serve as a selection parameter can be determined by genre, stylistic, temporal, anthropocentric, professional, communicative, geographical or social nature of linguistic diversity. Examples of these types of text corpora are also presented. The article presents terminological equivalents of corpus names by the type of text data in Ukrainian and English.

Key words: *general corpus, specialized corpus, text data, corpus type, typological characteristics.*

Вступ. В цілому, існує велика кількість різних типів корпусів, які науковці виокремлюють в залежності від поставлених цілей та класифікуючих ознак. Таке різноманіття спричинене потребами лінгвістичних досліджень. Спроби описати різні типи корпусів робили багато дослідників, як, наприклад, Демська-Кульчицька О. [1], Жуковська В. В. [2], Риков В. В. [3], Баранов А. Н. [4], Карпіловська Н. Є. [5], Кеннеді Ґ. та Синклер Дж. [6; 7], Деш Н. [8] До уваги бралися різні параметри та виділялися певні опозиції типів, як, наприклад, повнотекстові – фрагментарні, дослідницькі – ілюстративні, реєструвальні – інтерпретаційні, динамічні – статичні, діахронні – синхронні, загальномовні – спеціалізовані, усні – писемні і т.д.

Класифікуючи існуючі корпуси за типами, науковці брали до уваги різні аспекти, які можна підсумувати наступними запитаннями:

- які текстові дані бралися до уваги?
- який метод використовувався при створенні корпусу?
- який спосіб фіксації текстових категорій у корпусі?
- чи додаться щось до отриманих даних?
- з якою метою використовують кінцевий продукт?

Кожен корпус – це свій набір відповідей на поставлені вище запитання. Наприклад, *Британський національний корпус* (*British National Corpus*) за типом текстових даних є загальномовним, змішаного типу (вміщає як писемні, так і усні тексти), мономовним та синхронічним; за методом створення – збалансованим, фрагментним; за способом фіксації – електронним; за наявністю додаткових даних – морфологічно анотованим; за метою використання – дослідницьким.

Парадигма потенційних типологічних характеристик корпусу досить велика, що зумовлено індивідуальним характером завдань, які ставить лінгвіст у процесі наукового пошуку.

Завданням цієї статті ставимо дослідити різноманіття корпусів, зважаючи на те які тестові дані бралися до уваги при їх укладанні. Окремою метою вважаємо за потрібне визначити англо-українські відповідники назв кожного типу корпусу та представити відомі приклади таких корпусів.

Парадигма типологічних характеристик корпусу за типом тестових даних

<table>
<tr><th colspan="4">ТИП ТЕКСТОВИХ ДАНИХ</th></tr>
<tr><td colspan="4">І. За ступенем спеціалізації тестових даних</td></tr>
<tr><td rowspan="8">Загальномовний корпус</td><td rowspan="8">Спеціальний корпус</td><td colspan="2">2.1 за жанровою природою мовної різноманітності (літературний/фольклорний корпус, корпус текстів драматургії/ газетних статей/прози)</td></tr>
<tr><td colspan="2">2.2 за стилістичною природою мовної різноманітності (корпус текстів публіцистичного/ художнього/наукового стилю)</td></tr>
<tr><td colspan="2">2.3 за часовою природою мовної різноманітності (історичний корпус, корпус сучасної мови)</td></tr>
<tr><td colspan="2">2.4 за антропоцентричною природою мовної різноманітності (корпус учнівського мовлення, корпус мовлення підлітків/дітей/дорослих, авторський корпус)</td></tr>
<tr><td colspan="2">2.5 за професійною спеціалізацією мовної різноманітності (корпус політичного/економічного/комп’ютерного дискурсу)</td></tr>
<tr><td colspan="2">2.6 за комунікативною природою мовної різноманітності (корпус текстів-оригіналів, корпус перекладів, корпус діалогічного/монологічного мовлення, корпус промов)</td></tr>
<tr><td colspan="2">2.7 за географічною природою мовної різноманітності (діалектний корпус)</td></tr>
<tr><td colspan="2">2.8 за соціальною природою мовної різноманітності (корпус жаргону/арго/сленгу)</td></tr>
<tr><td colspan="4">ІІ. За формальною природою текстових даних</td></tr>
<tr><td colspan="2">3. Корпус писемного мовлення</td><td>4. Корпус усного мовлення</td><td>5. Змішаний корпус</td></tr>
<tr><td colspan="4">ІІІ. За мовним параметром текстових даних</td></tr>
<tr><td colspan="2">6. Одномовний корпус</td><td>7. Двомовний корпус</td><td>8. Багатомовний корпус</td></tr>
</table>

І. Класифікація корпусів за ступенем спеціалізації мовних даних, що входять до нього. За ступенем спеціалізації тестових даних, репрезентованих у корпусі виділяються *загальномовні* (*general corpus*) та *спеціальні* (*special/specialized corpus*) корпуси. Більшість науковців протиставляють корпуси, що стосуються усієї мови (часто мови певного періоду), корпусам, які представляють якусь мовну різноманітність (жанрову, стилістичну, певної вікової чи соціальної групи, письменника або вченого.). Представимо узагальнені дані щодо класифікації корпусів, беручи до уваги ступінь спеціалізації мовних даних, які входять до нього, у вигляді таблиці.

1. Загальньномовні корпуси (*general/reference corpus*) відображають усю різноманітність мовної діяльності. Вони слугують основою для опису всієї мови або якоїсь її різноманітності. Г. Кеннеді, у своїй, на сьогодні уже класичній у корпусному мовознавстві, праці «Вступ до корпусної лінгвістики» [9] вважає загальномовний (національний) корпус репрезентацією національної мовної системи та її реалізації. Для більшості мов світу уже створені національні лінгвістичні корпуси. *Британський національний корпус* (*British National Corpus*) [10], наприклад, покликаний повністю репрезентувати британський варіант сучасної англійської мови, а *Американський національний корпус* (*American National Corpus*) [11] – американський, *Корпус української мови* [12] – показує різні аспекти української мови, *Національний корпус російської мови* (*Национальный корпус русского языка*) [13] – показує різні аспекти російської мови. Загальномовний корпус представляє мову на певному етапі (чи етапах) її існування у всій різноманітності жанрів, стилів, територіальних і соціальних варіантів. Загальномовні корпуси збільшуються з часом, поповнюючись новими даними з новіших текстових зразків. Вони зазвичай великі за обсягом, вміщують велику кількість різноманітних текстів, які найповніше репрезентують мову чи її різноманітність. Іншим прикладом загальномовного корпусу є *Банк англійської мови* (*Bank of English*).

2. Спеціальні корпуси (*special/specialized corpus*) відображають існування деякого мовного чи культурного явища у суспільній мовній практиці, побудовані ad hoc (для спеціальних цілей), наприклад, корпус прислів’їв, чи корпус політичних метафор у публіцистиці; або такі, що створюються для вирішення спеціального завдання, як, наприклад, для налаштування системи автоматичного перекладу [14]. Спеціальні корпуси текстів зазвичай збалансовані, невеликі за розміром, підпорядковані одному дослідницькому завданню та призначені для використання переважно з метою, яку ставить перед собою кожен окремий дослідник чи група науковців. Обмежень щодо ступеню та різноманітності спеціалізації не існує, але в її межах висуваються чіткі параметри щодо вибору текстів. Наприклад, корпус може обмежуватися за часовою природою мовної різноманітності, складючись з текстів певного часового періоду, або соціальним характером, як, наприклад, з текстами розмов у книжковому магазині, чи певного соціального класу.

У межах групи спеціальних корпусів можна виділити підгрупи корпусів, текстові дані яких репрезентують мовну різноманітність різної природи:

2.1 За жанровою природою мовної різноманітності спеціальні корпуси можуть підрозподілятися на літературні, фольклорні, драматургічні, газетні та ін. корпуси [15]. Прикладами такого корпусу може слугувати *Корпус прози століття* (*The Century of Prose Corpus)*, у якому зібрано прозові тексти трактатів, написаних у проміжку часу між 1640 та 1740, що складають 1,1 мільйони слів., *Комп’ютерний корпус текстів російських газет кінця XX століття* (*Компьютерный корпус текстов русских газет конца XX века*) [16], *Корпус текстів журналу Time* (*Time magazine corpus*) [17].

2.2 За стилістичною природою мовної різноманітності спеціальні корпуси можуть містити тексти різних стилів: корпуси публіцистичного стилю наприклад, *Ростокський Історичний Корпус Англійських газет від 1700 і дотепер (The Rostock Historical English Newspaper Corpus from 1700 to today*), художнього – *Корпус Німецькомовної художньої літератури* (*Corpus of German-Language Fiction*), науково-популярного чи наукового, як наприклад, *Корпус англомовних наукових статей* (*SciCorp)*.

2.3 За часовою природою мовної різноманітності розрізняють *історичні корпуси (historical corpus)* та *корпуси сучасної мови (corpora of present-day language).*

Історичні корпуси (historical corpus) зорієнтовані на вивчення та аналіз еволюційних процесів у конкретній мові. Такий тип корпусів укладається переважно на матеріалі текстів однієї мови, відібраних у різні часові проміжки з метою вивчення історичної динаміки мовних змін [2, с 151]. Прикладами таких корпусів можуть слугувати підкорпуси *Національного кор-*

пусу російської мови: *Древнерусский подкорпус* [18], який включає в себе оригінальні давньоруські твори, виконані перекладі з грецької та пам'ятки південнослов'янського походження, переписані на Русі, та *Церковнославянский корпус* [19]. Для англійської мови створено *Пеннський корпус історичної англійської мови* (*Penn Corpora of Historical English)* [20], який складається з корпусів мовлення різних періодів: середньоанглійського – *Penn-Helsinki Parsed Corpus of Middle English* (*PPCME2*), ранньоанглійського – *Penn-Helsinki Parsed Corpus of Early Modern English* (*PPCEME*), та сучасного – *Penn Parsed Corpus of Modern British English* (*PPCMBE*), до яких увійшли повні тексти та текстові зразки Британської англомовної прози від найдавніших пам'яток до Першої світової війни.

Історичні корпуси за принципом підбору текстів до корпусу поділяються на *синхронні/синхронічні* та *діахронні/діахронічні*. Синхронні історичні корпуси базуються на текстах конкретних історичних періодів, наприклад, *Корпус прози століття (The Century of Prose Corpus)* обмежується текстами 1680-1780 рр., *Лемпетерський корпус трактатів ранньоновоанглійського періоду* (*Lampeter Corpus of Early Modern English Tracts*) [21] – спеціалізований історичний корпус писемного мовлення, до якого входять 120 трактатів, написаних у проміжку часу між 1640 та 1740, що складають 1,1 мільйони слів. Діахронні історичні корпуси охоплюють довші часові відрізки, як, наприклад, *Гельсінський корпус англійських текстів (Helsinki Corpus of English Texts)*, що включає тексти за десять століть (730–1710 рр.) [22].

2.4 За антропоцентричною природою мовної різноманітності виділяють *корпуси учнівського мовлення*, *корпуси мовлення підлітків/дітей/дорослих*, *авторські корпуси* і т.д.

Корпуси учнівського мовлення або *учнівські корпуси* (*learner corpora)* зазвичай укладаються з усних і/або писемних текстів, спродукованих особами, що вивчають мову як іноземну. Тут англійський термін learner, перекладається лексемою «учнівський», похідною від іменника «учень» зі значенням «той, хто вчиться, вивчає щось», тобто під цим прикметником слід розуміти людину, яка навчається, незалежно від віку. Такі корпуси почали створюватися ще в кінці 80-х на початку 90-х років XX століття. Під ним мається на увазі електронний корпус текстів для групи осіб, які вивчають іноземні мови. Основною метою організації учнівських корпусів є їх аналіз на предмет виявлення способів і ефективності освоєння мови. Такі корпуси можуть бути використані для лінгвістичного аналізу на предмет виявлення лексичних чи синтаксичних помилок при вивченні іноземної мови. Цей підхід допомагає встановити частотність тих чи інших типів мовних помилок, характерні контексти, що є необхідним для методичних прийомів для подальшої корекції у вивченні мови. Учнівські корпуси дуже поширені в Азії та Європі. Найбільш відомим є *Міжнародний корпус учнівської англійської мови* (*International Corpus of Learner English* (ICLE)). В основному цей корпус використовується для дискурсивного і статистичного аналізу вокабуляру учнів.

За типом можна виділити *комерційні учнівські корпуси* (*Лонгманський корпус учнівської англійської мови* (*Longman Learners' Corpus)* та *Кеймбріджський корпус учнівської англійської мови (Cambridge Learner Corpus*)), що ініціюються видавничими компаніями та *академічні* (*International Corpus of Learner English (ICLE), Санкт-Петербургский учебный корпус текстов школьников* (SPbEFLLC)) *корпуси*, які укладаються навчальними закладами.

Прикладами спеціальних корпусів, у яких за основу покладені тексти, продуковані певними групами людей є *Корпус мовлення лондонських тінейнджерів* (*Corpus of London Teenagers*) [23], *Корпус усного мовлення дітей* (*CHILDES*).

До *авторських корпусів* відносимо, наприклад, *Словник-корпус мови О.С. Грибоєдова* (*Словарь-корпус языка А.С. Грибоедова*).

2.5 За професійною спеціалізацією мовної різноманітності

Корпус мови для спеціальних цілей (sublanguage/specialized corpus). Цей корпус використовується для потреб в освіті та навчанні та для дослідження мовних змін у певній конкретній області знань. Корпус мови для спеціальних цілей широко застосовується для викладання іноземних мов, до прикладу для учнів, які мають в цьому потребу для їх освіти, професійної підготовки або роботи. Прикладами такого корпусу є: *Корпус мовлення авіадиспетчерів* (*Air Traffic Control Speech corpus*) [24], який складається з 50 сеансів моделюючих ситуацій у реальному часі; *Корпус професійного мовлення американської англійської мови* (*Corpus of Professional Spoken American English* (CPSA)) складається з транскриптів комунікативних ситуацій з академічної та політичної професійних галузей.

2.6 За комунікативною природою мовної різноманітності

Корпуси текстів-оригіналів (*non-translation corpus*) протиставляються *корпусам текстів-перекладів (translation corpus)* [25].

2.7 За географічною природою мовної різноманітності

Діалектний корпус текстів підвищує ймовірність збереження унікального матеріалу та створює можливість для вільнішого доступу до первинного діалектного матеріалу. Наприклад, *Корпус глобальної Web-базованої англійської мови* (*The Corpus of Global Web-Based English (GloWbE))* [26], може надати інформацію про різницю між 20 різними діалектами англійської мови.

2.8 За соціальною природою мовної різноманітності

До соціальних діалектів, за Л. Ставицькою, зараховують арго (особлива мова певної відокремленої професійної чи соціальної групи), жаргон (напіввідкрита лексико-фразеологічна підсистема, яку застосовують із метою відособлення від решти мовної спільноти) та сленг (різновид розмовної мови, яку суспільство оцінює як підкреслено неофіційну). Саме ці тексти у *соціолектних корпусах* є його репрезентативною ознакою, адже сповна відображають багатство та повну сучасну мовну картину світу кожного народу.

ІІ. За формальною природою текстових даних виділяються *корпуси усного мовлення*, *корпуси писемного мовлення* та *корпуси змішаного типу*.

4. Корпуси усного мовлення (speech corpus) включають тексти реальних усних комунікативних ситуацій. Це спеціальні колекції ретельно відібраних текстових уривків (слів, фраз, речень), вимовлених численними мовцями за різних акустичних умов. (Teubert 2007: 126). Зважаючи на трудомісткість і напруженість збору усних даних порівняно із писемними, ці корпуси є значно меншими за обсягом. Серед корпусів усного мовлення назвемо *Лондонсько-Лундський корпус усного мов-*

лення (*London-Lund Corpus of Spoken English (LLC)*); *Ланкастерський корпус усного мовлення* (*Lancaster/IBM Spoken English Corpus (SEC)*); *Кеймбріджсько-Нотінгемський корпус ділової англійської мови* (*Cambridge and Nottingham Spoken Business English Corpus (CANBEC))* – зібрання зразків усного мовлення, записаних на великій кількості зустрічей, які проходили у Великобританії у різноманітних комунікативних ситуаціях (наприклад, випадкові зустрічі, спілкування, пошук інформації, обговорення); *Корпус професійного усного мовлення американського варіанту англійської мови* (*Corpus of Professional Spoken American English (CPSA))* складається з транскриптів комунікативних ситуацій з академічної та політичної професійних галузей; *Мічіганський корпус академічного усного мовлення* (*Michigan Corpus of Academic Spoken English (MICASE))* – містить приблизно 1,7 млн. слововживань (близько 200 годин записів) сучасного усного університетського мовлення, що було записано в Мічиганському університеті; *C-ORAL-ROM* – багатомовний корпус спонтанного мовлення основних романських мов (французької, італійської, португальської та іспанської) на 1,2 млн. слів; *Корейський корпус усного мовлення* (*Korean Speech Corpus*); *Санта Барбарський корпус американського усного мовлення* (*Santa Barbara Corpus of American Spoken English*). Якщо враховувати класифікацію корпусів за способом фіксації текстових категорій у тексті (див. далі), то корпуси усного мовлення зазвичай представляються у вигляді корпусів транскрибованого мовлення, аудіо та відео корпусів. Термін ***усний корпус (spoken corpus)*** є технічним різновидом терміну *корпус усного мовлення* (*speech corpus*), оскільки позначає такий корпус, у якому зібрані зразки усного мовлення теговані з допомогою різних типів транскипції (наприклад, орфографічної, фонетичної).

У ***корпусах писемного мовлення*** (**3**) (*written corpus*) усний варіант реалізації мовної системи не представлений. У такому корпусі містяться лише зразки текстів, відібраних з різноманітних писемних, друкованих, або публікованих джерел. Найпершим сучасним корпусом писемного англійського мовлення був *Браунський корпус американського варіанту англійської мови* (*Brown University Standard Corpus of Present-Day American English)*, услід за яким було укладено низку подібних корпусів: *Австралійський корпус сучасної англійської мови* (*Australian Corpus of English*), *Колпахурський корпус англійської мови Індії* (*Kolhapur Corpus of Indian English*), *Корпус бангладешської англійської мови, створений студентами Массачусетського університету* (*MIT Bangla Corpus*). Це приклади корпусів, тексти до яких взяті з друкованого матеріалу. Якщо враховувати класифікацію корпусів за способом фіксації текстових категорій у тексті (див. далі), то корпуси писемного мовлення можуть бути, крім друкованих, ще й у електронній формі.

5. Змішаними корпусами за формальною природою текстових даних зазвичай бувають національні корпуси, які представляють існування мови в певний період часу (*Британський національний корпус* (*British National Corpus*), *Національний корпус російської мови* (*Национальный корпус русского языка*)).

III. За мовним параметром текстових даних розрізняють такі групи корпусів: *одномовні (monolingual), двомовні корпуси (bilingual), багатомовні корпуси (multilingual).*

6. Одномовні корпуси *(monolingual corpus)* складаються з текстових даних однією мовою. В залежності від спеціалізації такі корпуси можуть бути як загальномовними, так і спеціальними, за формальними ознаками – писемними, усними або змішаного типу. За методом створення та структурою одномовні корпуси можуть бути ***порівнянного типу*** – включати в себе підкорпуси різних діалектів чи варіантів однієї мови. Прикладом одномовного порівнянного корпусу може слугувати *Корпус міжнародної англійської мови* (*International Corpus of English – ICE*), проект, у якому вже зібрано 13 одномільйонних підкорпусів національно- та регіонально-особливих англомовних текстів, наприклад, канадійської, індійської, ірландської та ін. англійської мови. Навіть, коли говорити про *Британський національний корпус* (*British National Corpus*), до певної міри, його теж можна вважати одномовним корпусом порівнянного типу, адже він містить у собі підкорпуси різних мовленнєвих різноманітностей (регіональних, формальних, соціальних) англійської мови.

8. Багатомовні корпуси *(multilingual corpus)*, у загальному значенні, включають в себе тексти декількома різними мовами. У вузькому сенсі, корпус називають багатомовним, якщо до його складу входять тексти більше, ніж двома мовами. Корпуси текстів двома мовами називають ***двомовними корпусами*** (**7**) (*bilingual corpus*), наприклад, *Англо-норвезький паралельний корпус* (*English-Norwegian Parallel Corpus (ENPC)).*

Потрібно провести межу між двома типами багатомовних корпусів. Перший можна описати як невелике зібрання окремих одномовних корпусів, у тому значенні, що для кожної мови використовуються ті ж самі процедури та категорії, але кожен має абсолютно різні тексти цими декількома мовами. Цей тип корпусу зазвичай об'єднує тексти з однієї і тієї ж тематичної області, написані незалежно один від одного на двох або декількох мовах. Такі корпуси допомагають у роботі з термінологією та часто використовуються перекладачами. Мова йдеться про ***багатомовний порівнянний корпус*** (*comparable corpus*), наприклад, *Корпус контрактного права* (*Aarhus corpus in contract law)* датського, французького та англійського контрактного права складається з набору трьох одномовних корпусів текстів з юриспрунденції, які не є перекладами тих же ж самих текстів. Ще одним прикладом такого корпусу можна вважати *C-ORAL-ROM* – багатомовний корпус спонтанного мовлення основних романських мов (французької, італійської, португальської та іспанської) на 1,2 млн. слів.

Іншим типом багатомовного корпусу є (і є найбільш цікавим для досліджень) ***паралельний корпус***. Цей корпус стосується одних і тих же текстів на різних мовах. Якщо брати до уваги напрямок перекладу, то виділяють *однонаправлені* (*unidirectional*) *паралельні корпуси*, як наприклад, з української на англійську чи з англійської на українську мови; *двонаправлені* (*bi-directional*) *паралельні корпуси*, які наприклад, включають як оригінальні тексти українською мовою та їх переклади англійською, так і оригінальні тексти англійською та їх переклади українською; *різнонаправлені* (*multi-directional*) *паралельні корпуси* – корпуси, до яких, наприклад, увійшли оригінальні тексти українською мовою та їх переклади англійською, німецькою та французькою. До останньої категорії також можна віднести тексти, які продукуються одночасно декількома мовами (McEnery & Xiao 2007).

Багатомовні паралельні корпуси створюються відділами з комунікацій у багатомовних організаціях, як, наприклад, ООН, НАТО, ЄС та офіційно двомовних країнах, до прикладу – у Канаді. На даний час у наявності є декілька розмічених паралельних корпусів, а ті, що існують, зазвичай двомовні, а не багатомовні, як, наприклад, *Канадський корпус звітів парламенту (Canadian Hansard Corpus),* який складається з паралельних текстів французькою та англійською мовами, але

включає в себе обмежену кількість типів текстів (протоколи засідань канадського парламенту). Проекти, фінансовані ЄС (*Multilingual Aligned Annotated Corpus* (*CRATER*) та *Multilingual Text Tools and Corpora* (*MULTEXT*)) мають за свою мету створити унікальний багатомовний паралельний корпус. *Корпус паралельних текстів Європарл* (*Europarl parallel corpus*) включає в себе тексти засідань Європейського парламенту 21 європейською мовами.

Отже, аналіз типологічних характеристик корпусів текстів з огляду на те, які текстові дані бралися до уваги при його укладанні, зокрема за ступенем їх спеціалізації, формальною природою та за мовним параметром виявив, що парадигму за параметром «ступінь спеціалізації текстових даних» складають загальномовні та спеціалізовані корпуси; У свою чергу, у групі спеціалізованих корпусів тип текстових даних, які визначають назву корпусу, до якого вони входять та слугують параметром відбору, може визначатися жанровою, стилістичною, часовою, антропоцентричною, професійною, комунікативною, географічною чи соціальною природою мовної різноманітності. У подальших розвідках щодо типологічних характеристик корпусів текстів плануємо класифікувати корпуси з врахуванням й інших аспектів, а саме особливостей дизайну та структури корпусу, способу фіксації та індексації текстових даних у корпусі, а також особливостей використання корпусу.

Література:

1. Демська-Кульчицька О. Дещо про класифікацію текстових корпусів. *Наукові записки. Серія: Мовознавство.* 2004. 1 (11). С. 153–157.

2. Жуковська В. В. Ресурси корпусної лінгвістики у дослідженні історичної динаміки мови. *Слово і речення: синтактика, семантика, прагматика: матер. Міжнар. наук. конф.* / М-во осв. і науки України; Київ. ун-т ім. Б. Грінченка. – Київ: Київ. ун-т ім. Б. Грінченка, 2013. С. 151–156.

3. Рыков В. В. Прагматически ориентированный корпус текстов. *Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог – 99».* (Москва – Таруса). Москва, 1999. URL: http://rykov-cl.narod.ru/t.html (дата доступа 20.12.2015). Название с экрана.

4. Баранов А. Н. Введение в прикладную лингвистику [Текст]. Москва: УРСС Эдиториал, 2001. 358 с.

5. Карпіловська Н.Є. Вступ до комп'ютерної лінгвістики [Текст]. Донецьк: Юго-Восток, 2003. 183 с.

6. Sinclair J. M. Corpus typology [Text]. *A framework for classification. Studies in anglistics.* Stockholm: Almqvist & Wiksell. 1995. P. 17–33.

7. Sinclair J. M. Corpus Typology Draft. 1996. URL : http://www.ilc.cnr.it/EAGLES/typology/typology.html (access date 20.12.2015). – Title from the screen.

8. Dash, Niladri Sekhar. Corpus Linguistics and Language Technology : With Reference to Indian Languages. New Delhi : Mittal Publications, 2005.

9. Kennedy G. Introduction to Corpus Linguistics. London-New-York: Longman, 1998. 309 p.

10. British National Corpus. *Oxford Text Archive* / IT Services. University of Oxford, 2009. URL : http://www.natcorp.ox.ac.uk/ (access date 20.12.2015). – Title from the screen.

11. American National Corpus. *American National Corpus Project*. 2009. URL : http://www.anc.org/ (access date 20.12.2015). – Title from the screen.

12. КОРПУС УКРАЇНСЬКОЇ МОВИ / Н. П. Дарчук (керівник проекту), В. М. Сорокін (програміст), О. Б. Сірук та інш. ; лабораторія комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка. *MOVA.info (ЛІНГВІСТИЧНИЙ ПОРТАЛ. К. : Київський національний університет імені Тараса Шевченка)*, 2003. URL : http://www.mova.info/corpus.aspx?l1=209 (дата доступу 20.12.2015). Назва з екрану.

13. Национальный корпус русского языка / Национальный корпус русского языка. 2003. URL : http://www.ruscorpora.ru/ (дата доступа 20.12.2015). Название с экрана.

14. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн. Санкт-Петербург: СПбГУ. РИО. Филологический факультет. 2013. С. 14.

15. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн. Санкт-Петербург: СПбГУ. РИО. Филологический факультет. 2013. С. 17.

16. Компьютерный корпус текстов русских газет конца XX-ого века / Лаборатория общей и компьютерной лексикологии и лексикографии. *Сайт филологического факультета МГУ имени М. В. Ломоносова.* Москва, 2001. URL : http://www.philol.msu.ru/~lex/corpus/ (дата доступа 20.12.2015). Название с экрана.

17. TIME Magazine Corpus. *corpus.byu.edu* / Mark Davies, BYU (Google Scholar), 2014?. URL : http://corpus.byu.edu/time/ (access date 20.12.2015). Title from the screen.

18. Древнерусский корпус. *Национальный корпус русского языка* / Национальный корпус русского языка. 2003. URL : http://www.ruscorpora.ru/search-old_rus.html (дата доступа 20.12.2015). Название с экрана.

19. Церковнославянский корпус. *Национальный корпус русского языка* / Национальный корпус русского языка. 2003. URL : http://www.ruscorpora.ru/search-orthlib.html (дата доступа 20.12.2015). Название с экрана

20. Penn Parsed Corpora of Historical English. *Department of Linguistics*. 2015?. URL : http://www.ling.upenn.edu/hist-corpora/ (access date 20.12.2015). Title from the screen.

21. The Lampeter Corpus of Early Modern English Tracts. *University of Oxford Text Archive* / University of Oxford. 2015?. URL : http://ota.ox.ac.uk/headers/2400.xml (access date 20.12.2015). Title from the screen.

22. Xiao R. Well-known and Influential Corpora. *Corpus Linguistics. An International Handbook*. Edited by A. Lüdeling, M. Kytö. 2008. Volume 1. P. 401.

23. The Bergen Corpus of London Teenage Language / A Language Research Project at the University of Bergen. *British National Corpus*. 2009?. URL : http://clu.uni.no/icame/colt/ (access date 20.12.2015). Title from the screen.

24. ATCOSIM: Air Traffic Control Simulation Speech Corpus / Signal Processing and Speech Communication Laboratory. EEC and Graz University of Technology. 2008?. URL : https://www.spsc.tugraz.at/tools/atcosim (access date 20.12.2015). Title from the screen.

25. McEnery, AM & Xiao, 2007, 'Parallel and comparable corpora: What are they up to?'. in G James & G Anderman (eds), Incorporating Corpora: Translation and the Linguist. Translating Europe, Multilingual Matters, Clevedon, UK, 2007. ISBN 978-1-85359-986-6. URL : http://eprints.lancs.ac.uk/59/1/corpora_and_translation.pdf (access date 20.12.2015).

26. Corpus of Global Web-Based English (GloWbE). *corpus.byu.edu* / Mark Davies, BYU (Google Scholar). 2014?. URL : http://corpus.byu.edu/glowbe/ (access date 20.12.2015). Title from the screen.